改进粒子群联合禁忌搜索的特征选择算法

张震[1]，魏鹏[1]，李玉峰[1]，兰巨龙[1]，徐萍[2]，陈博[1]
（1. 国家数字交换系统工程技术研究中心，河南 郑州 450002；
2. 中国人民解放军战略支援部队信息工程大学，河南 郑州 450002）

摘 要：针对入侵检测中数据特征维度高的问题，提出了改进粒子群联合禁忌搜索(IPSO-TS)的特征选择算法。采用遗传算子对粒子群算法进行了改进，得到了特征选择初始最优解；对该解进行禁忌搜索(TS)得到了特征子集的全局优化解。基于KDD CUP 99数据集的实验结果表明，相较遗传算子整合粒子群算法(CMPSO)、粒子群算法(PSO)和粒子群联合禁忌算法，IPSO-TS 减少了至少 29.2%的特征，缩短了至少 15%的平均检测时间，提高了至少 2.96%的平均分类准确率。

关键词：入侵检测；特征选择；粒子群；遗传算法；禁忌搜索
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000−436x.2018287

Feature selection algorithm based on improved particle swarm joint taboo search

ZHANG Zhen[1], WEI Peng[1], LI Yufeng[1], LAN Julong[1], XU Ping[2], CHEN Bo[1]
1. National Digital Switching System Engineering and Technological Research and Development Center, Zhengzhou 450002, China
2. Information Engineering University, Zhengzhou 450002, China

Abstract: To solve the problem of high data feature dimensionality in intrusion detection, a feature selection algorithm based on improved particle swarm optimization taboo search (IPSO-TS) was proposed. The genetic algorithm was used to improve the particle swarm optimization, and the initial optimal solution of feature selection was obtained. A taboo search (TS) algorithm was used for initial optimal solution to obtain the global optimal solution of the feature subset. Compared with genetic algorithm integrated particle swarm optimization (CMPSO), particle swarm optimization (PSO) and PSO-TS algorithms, experimental results based on the KDD CUP 99 dataset show that the method reduces the features by about 29.2% , shortens about 15% of the average detection time, and increases about 2.96% of the average classification accuracy.

Key words: intrusion detection, feature selection, particle swarm optimization, genetic algorithm, taboo search

1 引言

入侵检测技术利用数据挖掘、机器学习等方法对审计日志、安全日志或者网络上获取的信息进行分析处理，实现对入侵行为的检测[1]。瞬息万变的海量网络数据使入侵检测系统（IDS，intrusion detection system）面临巨大挑战，如检测速度低、检测效果差、计算负荷大等。其中，检测速度是入

收稿日期：2018−01−09；**修回日期**：2018−05−14
基金项目：国家重点研究发展计划基金资助项目（No.2017YFB0803201）；国家自然科学基金资助项目（No.61502528）；网络空间安全专项课题基金资助项目（No.2017YFB0803204）；上海市科学技术委员会科研计划课题基金资助项目（No.16DZ1120503）
Foundation Items: The National Key Research and Development Program (No.2017YFB0803201), The National Natural Science Foundation of China (No.61502528), The Network Space Security Special Project (No.2017YFB0803204), The Shanghai Science and Technology Commission Research Project (No.16DZ1120503)

侵检测系统实时性要求的重要指标，如何在保证检测准确率的前提下提升入侵检测速度，成为当前入侵检测的研究热点。

特征选择是入侵检测中数据预处理的关键环节，能够在入侵检测数据集中筛选出对分类器的分类性能影响最重要的一组特征[2]。良好的特征选择算法能够降低网络数据的维度，减轻入侵检测系统的计算负荷，提高入侵检测系统的检测速度。本文针对原始数据集数据维度高的问题，提出了一种改进粒子群联合禁忌搜索（IPSO-TS, improved particle swarm optimization taboo search）的特征选择方法，通过降低入侵数据的维度，有效地提升了入侵检测效率。

2　研究现状

特征选择算法根据是否独立于分类算法可分为 Filter 方法和 Wrapper 方法。Filter 方法独立于分类算法，首先衡量每个特征的重要性，然后根据重要性对所有特征进行排序，最后将前 N 个重要的特征作为特征子集。这种方法的优点是计算量较低，可以有效去除噪声特征，缺点是忽略了选择出的特征子集整体对分类效果的影响程度[3]。Wrapper 方法通过分类算法的准确率评估特征子集的优劣，优点是考虑了整个特征子集对分类效果的影响，分类准确率较高，缺点是计算量较大[4]。从分类性能方面考虑，Wrapper 方法比 Filter 方法更适合于网络数据的特征选择，故本文采用 Wrapper 方法。

搜索机制是 Wrapper 方法的核心，用来生成候选特征子集[5]。考虑到对庞大而复杂的搜索空间中的所有可能特征子集进行穷举搜索很难实现，粒子群算法（PSO, particle swarm optimization）为代表的演化计算[6]具有自然进化的属性，在特征选择中有利于优化搜索机制，提高搜索效率，故常被用来寻优特征子集。近年来，有许多结合改进粒子群进行特征选择的方法，本文从优化粒子群算法参数、优化粒子群表示方法、优化离散型或连续型特征、优化评价指标等方面引用相关参考文献进行分析。董跃华等[7]提出一种自适应粒子群联合禁忌搜索的特征选择算法，在 PSO 搜索过程中，对每一代全局优化粒子进行禁忌搜索，有利于保持粒子活力，增强搜索优化解的能力，但该方法中粒子群缺乏多样性且对每一代全局优化粒子进行禁忌搜索会极大增加算法的时间复杂度。Tran 等[8]提出了基于势粒子群优化的特征选择方法，设计了一种新的粒子群表示方法，通过选取切点离散化原始特征，然后筛选出性能较好的优化特征子集，该方法仅适用于离散型特征，无法筛选出连续型特征中的优化特征。Nguyen 等[9]提出了一种遗传算子优化 PSO 的连续型特征选择算法，利用交叉、变异等遗传算子提高 PSO 的搜索能力，但该方法受群落初始分布影响较大，且遗传算子中交叉、变异概率固定不变，不利于搜寻到全局优化解。翟俊海等[10]提出了一种将粗糙集相对分类信息熵和粒子群算法相结合的特征选择方法，将分类信息熵作为适应度函数，结合改进粒子群算法得到优化的特征子集，但该方法的适应度评价指标结构较为单一，不利于选出分类性能较好的特征子集。

综上所述，目前大多数特征选择算法都有各自的局限性，例如，文献[8]中方法无法筛选出连续型特征中的优化特征；文献[9]中算法寻优效果受群落初始分布影响较大；文献[10]中方法的评价指标结构较为单一。本文以特征维度和分类准确率的优化配置为目标设计出评价标准，基于该标准提出了一种遗传算子改进粒子群联合禁忌搜索的特征选择（IPSO-TS）算法，通过遗传算子优化粒子群算法寻优特征子集，以期望生成更好的初始最优解。在此基础上，联合禁忌搜索（TS, taboo search）算法找到全局优化解。其中，粒子群算法具有快速收敛的特点；遗传算法增加了粒子群落多样性；TS 算法有利于跳过局部最优值，实现全局优化搜索，提高特征子集搜索效率。

3　IPSO-TS 特征选择算法

IPSO-TS 特征选择算法的流程如图 1 所示，其中，搜索模块应用本文提出的改进粒子群联合 TS 算法，实现搜寻候选特征子集的功能；评价模块应用本文提出的适应度函数，完成对候选特征子集的性能评价；判决模块根据判决条件选择数据集的最优特征子集，结束循环，并将该子集作为结果输出。

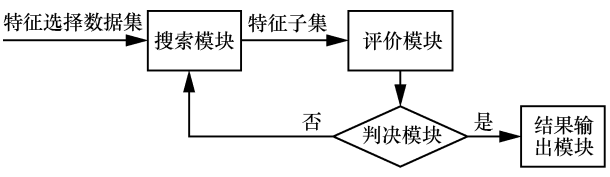

图 1　IPSO-TS 特征选择算法流程

针对图 1 中的评价模块，本文提出了一种新的适应度评价函数，如式(1)所示。

$$fvalue = alpha\left(1-\frac{fsnum}{afsnum}\right)+(1-alpha)accuracy \tag{1}$$

其中，*fsnum* 表示特征选择后的特征数量，*afsnum* 表示总特征数量，*accuracy* 表示分类准确率，*alpha* 和 (1–*alpha*)都是权重参数。本文的目标是在保障分类准确率的条件下，降低特征维数，因此，选择了较高的 *accuracy* 权重参数，即 *accuracy*=0.98，相应的 *alpha*=0.02。

式(1)中适应度函数与传统特征选择方法相比，引入了分类准确率 *accuracy* 和特征选择后的特征数量 *fsnum* 作为参数，能够更准确对特征子集性能进行评估。其中，分类准确率越高，特征维数减少的比例 $\left(1-\frac{fsnum}{afsnum}\right)$ 越大，则适应度值越大，表明提取的特征子集性能越好。

本文通过 K 最近邻(KNN，k-nearest neighbor)分类算法得到分类准确率 *accuracy*，如式(2)所示。

$$accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

其中， *TP* 和 *TN* 表示入侵检测中正确分类的正常数据和异常数据，*FP* 和 *FN* 表示入侵检测中错误分类的正常数据和异常数据。

3.1 改进的 PSO 算法

3.1.1 PSO 算法

PSO 算法是根据鸟群捕食行为设计的一种群智能算法，基本思想是通过粒子群中粒子之间的协作和信息共享来搜寻最优解。PSO 算法可以分为二进制 PSO 算法和连续 PSO 算法。连续 PSO 算法指为粒子赋值实数，二进制 PSO 算法则是为粒子赋值二进制数。由于连续 PSO 具有更广阔的应用前景，本文使用连续 PSO 算法进行特征选择，根据粒子个体极值和群体最优解进行寻优，寻优计算式如式(3)和式(4)所示。

$$\begin{aligned}vstep_{id}^{t+1}=&w_i vstep_{id}^{t}+c_1 rand_{i1}(pbest_{id}^{t}-x_{id}^{t})+\\&c_2 rand_{i2}(gbest_{d}^{t}-x_{id}^{t})\end{aligned} \tag{3}$$

$$x_{id}^{t+1}=x_{id}^{t}+vstep_{id}^{t+1} \tag{4}$$

其中，*t* 表示迭代次数，*i*=1,2,⋯,*m* (*m* 表示所有粒子个数)，*d*=1,2,⋯,*n* (*n* 表示粒子总维数)，$rand_{i1}$ 和 $rand_{i2}$ 为[0,1]之间的随机数，$pbest_{id}^{t}$ 表示粒子 *i* 的历史极值在第 *t* 次迭代中第 *d* 维最优位置，$gbest_{d}^{t}$ 表示群体最优粒子在第 *t* 次迭代中第 *d* 维最优位置，c_1 和 c_2 表示学习因子，x_{id}^{t} 表示粒子 *i* 在第 *t* 次迭代中第 *d* 维位置，$vstep_{id}^{t}$ 表示粒子 *i* 在第 *t* 次迭代中第 *d* 维移动速度。

刘杨等[11]研究了线性函数、凹凸函数递减等 4 种 PSO 算法惯性权重 *w* 后发现，多峰函数寻优中凸函数递减惯性权重法收敛速度最快，效果最好。因此本文的惯性权重取值如式(5)所示。

$$w_i=(w_{\max}-w_{\min})\left(1-\frac{iter}{maxiter}\right)^3+w_{\min} \tag{5}$$

其中，$w_{\max}$ 和 $w_{\min}$ 分别表示惯性权重的最大值和最小值，*iter* 表示迭代次数，*maxiter* 表示最大迭代次数。

在 PSO 算法中，c_1 体现了粒子对自身的学习能力，c_2 体现了粒子对群体的学习能力[12]，c_1 数值递减和 c_2 数值递增有利于发挥粒子初期探索和后期对群体的认知能力。因此本文学习因子 c_1 和 c_2 的赋值如式(6)和式(7)所示。

$$c_1=c_3-c_4\left(\frac{iter}{maxiter}\right) \tag{6}$$

$$c_2=c_4+c_3\left(\frac{iter}{maxiter}\right) \tag{7}$$

其中，c_3 和 c_4 是常数，且 $c_3>c_4$。

3.1.2 基于遗传算法对粒子群的改进

遗传算法是一种通过模拟自然进化过程寻找最优解的随机化搜索方法。遗传算法中的遗传算法交叉和突变对于 PSO 算法作用很大，可以增加粒子群落变化的多样性，产生出代表新解集的种群，克服粒子群易于陷入局部最优的问题。

为了避免粒子群过早收敛问题，本文中的粒子群算法每次迭代的时候对多对粒子进行交叉操作，利用基于适应度值的轮盘赌算法选择出多对母本粒子，将其进行交叉产生出多对下一代粒子，选择其中优良的下一代粒子替换群落中母本粒子和历史极值。其中交叉操作主要是将第 1 个母本粒子 $1\sim\lfloor\frac{L}{2}\rfloor$ 与第 2 个母本粒子 $L-\lfloor\frac{L}{2}\rfloor+1\sim L$ 维位置互相交换，其中，*L* 表示粒子总维数。每次迭代进行交叉操作粒子的个数 P_i 如式(8)所示。

$$P_i=\left\lfloor swarmsize-\left(\frac{iter}{maxiter}\right)^2(swarmsize-2)\right\rfloor \tag{8}$$

其中，*swarmsize* 表示所有粒子个数。根据粒子群的特性，遗传算法在搜索初期时粒子群的种类更加随机和多样，到了搜索后期，粒子群逐渐收敛，相似度增加，故交叉操作在前期作用大于后期作用。所以随着迭代次数增加，每次迭代的交叉操作粒子个数应该随之减少，且交叉概率应该随之降低，这样有利于减少计算成本。交叉概率更新如式(9)所示。

$$pcross = pc_{\max} - pc_{\min}\left(1-\frac{iter}{maxiter}\right)+pc_{\min} \quad (9)$$

其中，$pc_{\min}$ 和 $pc_{\max}$ 分别表示最小和最大交叉概率。

虽然交叉操作初期可以很好地探索搜索空间，但在粒子群后期收敛后作用较小[13]。变异操作和交叉操作恰好相反，在搜索初期效果不明显，但是在搜索后期有较好的搜索效果，于是引入变异操作以进一步提高空间搜索能力，且随着迭代次数增加变异概率随之提高，有利于进一步寻找到更优粒子。变异概率更新如式(10)所示，粒子中第 d 维位置的选择概率如式(11)所示。

$$pmutation = (pm_{\max} - pm_{\min})\frac{iter}{maxiter} + pm_{\min} \quad (10)$$

$$CR_d = \begin{cases} \dfrac{(gbest_d-\theta)^2}{1-\theta} & ,\theta < gbest_d \leqslant 1 \\ \dfrac{(\theta-gbest_d)^2}{\theta} & ,0 \leqslant gbest_d \leqslant \theta \end{cases} \quad (11)$$

其中，$pm_{\min}$ 和 $pm_{\max}$ 分别表示最小和最大变异概率，$gbest_d$ 表示全局最优粒子第 d 维位置，θ 表示选择阈值。当 *gbest* 第 d 维位置的值大于阈值 θ 时，若 $gbest_d$ 值越大，则 CR_d 越大，否则 CR_d 越小。当 *gbest* 第 d 维位置的值小于阈值 θ 时，若 $gbest_d$ 值越小，则 CR_d 越大，否则 CR_d 越小。然后生成[0,1]之间的随机数 r，若 $r>CR_d$，执行相应的变异操作，如式(12)所示，否则将新粒子第 d 维位置 $child_d$ 赋值为 $gbest_d$。

$$child_d = \begin{cases} \theta+(1-\theta)\dfrac{gbest_d}{\theta} & ,0 \leqslant gbest_d \leqslant \theta \\ \theta-\dfrac{(1-gbest_d)}{1-\theta}\theta & ,\theta < gbest_d \leqslant 1 \end{cases} \quad (12)$$

其中，若 $gbest_d$ 的值小于或等于选择阈值 θ，则将 $child_d$ 的值映射到[θ,1]之间的实数，否则将 $child_d$ 的值映射到(0,θ]之间的实数，这有利于改变特征的选择属性。

本文中改进 PSO 算法通过不断进化，后期粒子群逐渐收敛，粒子群相似度大大增加，这导致很难搜寻到更好的优化解。由于 TS 算法不仅可以接受优解，还可以接受劣解，比 PSO 算法能获得更优解，因此本文通过引入 TS 算法解决了该问题。

3.2　TS 算法

TS 算法是一种亚启发式随机搜索全局寻优算法，主要包含初始解、邻域函数、禁忌表属性、候选集藐视准则等基本参数[14]，以局部邻域搜索算法为基础，通过设置禁忌表停止已经进行过的操作或变换，再利用相应的藐视准则释放禁忌表中符合条件的元素[15]。

本文建立长度为 L 的禁忌表 (L 表示总特征个数)，设置迭代次数为$\lfloor\frac{L}{3}\rfloor$，将通过改进粒子群算法得到的初始最优解作为 TS 算法的初始解。改变初始解 2 个随机位置特征元素的值，保持其他特征元素的值不变，将该过程作为邻域函数：设初始最优解为 $\boldsymbol{R}=\{T_i,i=1,2,\cdots,L\}$，随机生成 2 个不同整数随机数 x 和 y，利用式(13)更新初始最优解中 T_x 和 T_y 的值，得到新的候选最优解。通过随机生成$\left\lfloor\frac{L}{2}\right\rfloor$组不同随机数对 x 和 y，对初始特征集进行变换，得到$\left\lfloor\frac{L}{2}\right\rfloor$组候选解当作初始候选解集合。

$$T_t = \begin{cases} \theta+(1-\theta)\dfrac{T_t}{\theta}, & 0 \leqslant T_t \leqslant \theta \\ \theta-\dfrac{(1-T_t)}{1-\theta}\theta, & \theta < T_t \leqslant 1 \end{cases} \quad (13)$$

其中，t 表示 x 或 y。

本文中禁忌表使用队列结构，每次迭代时，添加禁忌对象到队首，队列溢出时，删除排在队尾的禁忌元素。藐视准则要满足 2 个要求：1) 将要解禁的解必须优于当前解[16]；2) 为了避免陷入死循环，禁忌表队首的元素不能被解禁。

4　IPSO-TS 特征选择方法的 IDS 应用实例

IDS 流程如图 2 所示，其中，数据获取模块通过监控目标网络或系统来采集数据；数据预处理模块是对采集的数据进行数值化、标准化、特征选择等处理，本文 IPSO-TS 群算法用于选择入侵检测数据的特征子集；入侵检测模块利用训练数据来训练分类器，采用分类器对待检测的数据集进行分析，获得入侵检测的结果，通过检测指标评价结果，本文中的分类器由 KNN 算法构建；响应模块是 IDS 对检测结果的决策。

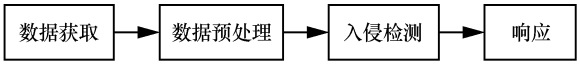

图 2 IDS 流程

在IPSO-TS算法中，若粒子位置元素大于阈值，表示选择该特征，否则放弃该特征。通过粒子群的移动产生不同位置向量，记录粒子搜索期间所有粒子历史极值和群落最优粒子，将收敛后的群落最优粒子当作 TS 算法的初始解。IPSO-TS 算法的特征选择方法基本流程如图 3 所示，具体实现步骤如下。

步骤 1 设置相关参数。改进粒子群最大循环次数 $maxiter_1$=50，禁忌搜索次数 $maxiter_2$=30，粒子群规模 *swarmsize*=30，最大适应度 *maxfit*=0.9999，$pc_{\min}$=0.7，$pc_{\max}$=0.8，$pm_{\min}$=0.02，$pm_{\max}$=0.1。设置惯性权重 *w*=0.7928，阈值 *threshold*=0.7 等推荐参数[4]。文中入侵检测数据集特征赋值为在[0,1]之间的不同实数，构成粒子位置向量，随机初始化粒子群位置和速度，求解适应度，初始化粒子群 *swarm*、粒子历史极值 *pbest*、群体最优解 *gbest* 及它们的适应度 *fswarm*、*fpbest*、*fgbest*。其中，选择阈值 *threshold* 设置为 0.7，目的是大概率筛选出 30%的特征子集；

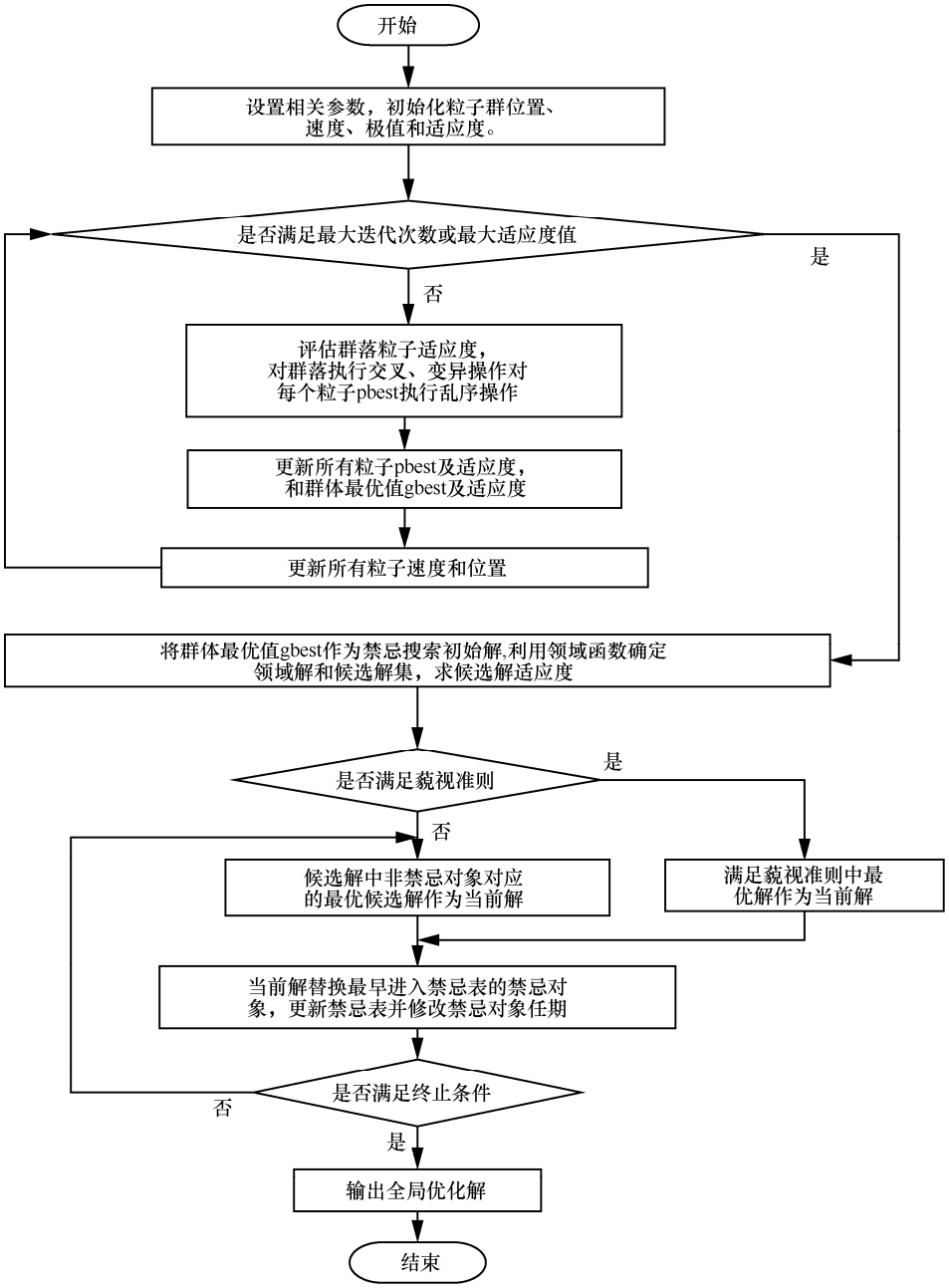

图 3 IPSO-TS 算法在 IDS 中应用流程

步骤 2　判断是否满足 $iter>maxiter_1$ 或 $fgbest>maxfit$，满足则执行步骤 6，否则执行步骤 3。

步骤 3　评估粒子群体适应度值，根据式(5)~式(7)更新惯性权重 w、学习因子 c_1、c_2，如式(9)和式(10)更新交叉、变异概率。第 *iter* 次迭代中，根据 3.1.2 节对群落中粒子对执行交叉操作、对所有 *pbest* 执行随机乱序操作以及对 *gbest* 执行突变操作分别得到各自后代粒子，若后代粒子更优，则将父母粒子替换为后代粒子。其中，对 *pbest* 执行随机乱序操作目的是通过随机打乱所有粒子历史极值位置的顺序，以找到更优的粒子个体极值，有利于避免粒子个体极值过快陷入局部最优解。

步骤 4　分别更新相应 *pbest*、*gbest* 及它们的适应度。

步骤 5　更新粒子群体的速度和位置，然后执行步骤 2。

步骤 6　根据 3.2 节进行禁忌搜索，将经过上述步骤得到的群体最优解 *gbest* 作为 TS 算法的初始解，利用初始解产生相应的邻域解，得到 $\left\lfloor \frac{L}{2} \right\rfloor$ 组不同解当作初始候选集合及其适应度。

步骤 7　判断候选解集中的解是否满足藐视准则，若满足则执行步骤 8，否则执行步骤 9。

步骤 8　将候选解中非禁忌对象对应的最优候选解作为当前解，然后执行步骤 10。

步骤 9　将满足藐视准则中最优解作为当前解，然后执行步骤 10。

步骤 10　将最早进入禁忌表的禁忌对象作为当前解，更新禁忌表并修改禁忌对象任期。

步骤 11　判断是否满足最大禁忌搜索次数或当前解适应度是否达到限定值，若满足则执行步骤 8，否则执行步骤 12。

步骤 12　输出全局优化解作为入侵检测数据集全局优化特征子集。

5　IDS 实验结果与分析

5.1　KDD CUP 99 数据预处理

本文实验中使用的是 KDD CUP 99 数据集，它是网络入侵检测领域的标准数据集[17]，选取其提供的 10%训练子集和测试子集（corrected）进行实验，其中，训练子集包含 49 万条网络连接记录，测试子集包含 31 万条网络连接记录。每个网络连接被标记为正常类型和异常类型，异常类型被分为 DoS（拒绝服务攻击）、Probe（扫描与探测）、R2L（未经授权的远程访问）、U2R（对本地超级用户的非法访问）等 4 类，共有 39 种异常类型。本文实验中分别针对 4 类异常类型数据和汇总数据，进行入侵攻击检测，样本数据集如表 1 所示。

数据集中每一条记录代表一条完整的会话，每条连接记录由 4 类特征集组成：基本特征集、内容特征集、基于时间特征集和基于主机特征集，共有 41 个特征，其中包含 3 个字符特征和 38 个数值特征，如表 2 所示。

表 1　**KDD CUP 99 样本数据集组成**

数据类型	10%训练数据集			测试数据集（corrected）		
	异常样本量	正常样本量	总样本量	异常样本量	正常样本量	总样本量
异常类型数据 DoS	391 458	91 848	483 306	229 853	45 941	275 794
异常类型数据 Probe	4 107	4 200	8 307	4 166	4 130	8 296
异常类型数据 R2L	1 126	1 130	2 256	16 189	10 112	26 231
异常类型数据 U2R	52	100	152	228	410	638
汇总数据	396 743	97 278	494 021	250 436	60 593	311 029

表 2　**KDD CUP 99 连接记录的特征集**

特征分类	特征名称
tcp 连接基本特征(9 个特征)	1) duration, 2) protocol_type, 3) service, 4) flag, 5) src_bytes, 6) dst_bytes, 7) land, 8) wrong_fragment, 9) urgent
tcp 连接的内容特征(13 个特征)	10) hot, 11) num_failed_logins, 12) logged_in, 13) num_compromised, 14) root_shell, 15) num_root, 16) su_attempted, 17) num_file_creations, 18) num_shells, 19) num_access_files, 20) num_outbound_cmds, 21) is_host_login, 22) is_guest_login
基于时间的网络流量统计特征(9 个特征)	23) count, 24) srv_count, 25) serror_rate, 26) srv_serror_rate, 27) rerror_rate, 28) srv_rerror_rate, 29) same_srv_rate, 30) diff_srv_rate, 31) srv_diff_host_rate
基于主机的网络流量统计特征(10 个特征)	32) dst_host_count, 33) dst_host_srv_count, 34) dst_host_same_srv_rate, 35) dst_host_diff_srv_rate, 36) dst_host_same_src_port_rate, 37) dst_host_srv_diff_host_rate, 38) dst_host_serror_rate, 39) dst_host_srv_serror_rate, 40) dst_host_rerror_rate, 41) dst_host_srv_rerror_rate

特征选择前需要对数据进行预处理，入侵检测中数据预处理主要分为字符特征数值化和数据归一化过程。字符特征数值化是指将符号特征映射到有序数字，例如 protocol_type 类型映射为 tcp = 1，协议类型映射为 udp = 2，协议类型映射为 icmp = 3。以同样的方式，将 service、flag 和数据集中的类别特征映射到有序数字。对经过字符特征数值化后的数据进行归一化处理，使特征数值处于相同数量级，目的是消除因特征数值范围不同造成的影响。本文通过式(14)进行数据归一化处理。

$$Y = \frac{Y_{\text{original}} - Y_{\min}}{Y_{\max} - Y_{\min}} \tag{14}$$

其中，在入侵检测数据集的同一维数据列中，Y_{original} 表示该列原始数值，$Y_{\min}$ 表示数据列中最小值，$Y_{\max}$ 表示数据列中最大值。

5.2 实验结果及分析

本文实验数据集主要分为检测样本数据集和特征选择数据集。入侵检测样本数据集采用表 1 中的训练和测试数据集。为了缩短寻优特征子集的时间，从表 1 的 DoS、Probe、R2L 和 U2R 的训练数据集和测试数据集中分别随机抽取 10%、20%、20%和 100%的数据作为各自特征选择数据集中的训练和测试数据，并将上述特征选择数据集合并作为汇总数据的特征选择数据集。

本文实验的计算机操作系统为 Windows 10，处理器为 Intel(R) Core(TM) i5-3230M 2.60 GHz，内存为 4 GB，测试环境为 Matlab 2014b。本文所提 IPSO-TS 算法从特征选择时间和特征选择维数，检测时间和分类准确率，漏报率和误报率这 6 个方面，对比了 PSO-TS 算法[7]、PSO 算法和 CMPSO（co-evolutionary particle swarm optimization）算法[9]，得到如下实验结果。

在特征选择数据集上比较 4 种特征算法得到特征子集的特征选择时间和特征选择维数，如表 3 所示。其中特征选择时间指对特征选择数据集进行特征选择的总用时，特征维数序号和表 2 相对应。

从表 3 中可以得到，通过与 PSO-TS 算法、CMPSO 算法和 PSO 算法得到的各个检测模型平均特征选择维数比较发现，本文所提 IPSO-TS 算法比其他特征选择方法减少了至少 29.2%的特征。本文方法的特征选择时间比其他特征选择算法要长，主要是由于 *pbest* 随机乱序过程和 TS 过程增加了特征选择时间。

在表 1 检测样本数据集上比较所有特征情况下和上述 4 种特征选择算法后的入侵检测时间和分类准确率，其中入侵检测时间是指利用 KNN 算法[18]对表 1 检测样本数据集中的测试数据集进行分类预测的总用时，K 取值为 5。实验结果如下表 4。

从表 4 中可以得到，通过与 PSO-TS 算法、CMPSO 算法和 PSO 算法得到的特征子集作为分类样本的平均分类准确率比较发现，本文方法比其他特征选择方法提高了至少 2.96%的平均分类准确率，缩短了至少 15%的平均检测时间；比所有特征情况提高了 0.27%的平均分类准确率，缩短了至少 42.32%的平均检测时间。其中，本文方法特征选择时间相比其他特征选择算法更长，该时间主要耗费

表 3　4 种特征选择算法的特征选择时间和特征选择维数

入侵检测模型	特征选择维数（特征个数/个）				特征选择时间/s			
	IPSO-TS	CMPSO	PSO-TS	PSO	IPSO-TS	CMPSO	PSO-TS	PSO
DoS	4(2,4,8,13)	5(3,25,28,34,35)	6(2,14,23,31,34,36)	10(1,3,4,5,6,11,12,16,19,20)	9845.27	9541.10	8836.66	7218.32
Probe	3(22,33,35)	4(23,31,38,41)	5(2,4,8,13,33)	8(2,23,25,31 32,35,36,39)	5128.18	4330.32	4635.58	3171.21
R2L	3(2,4,13)	3(5,10,15)	7(2,3,4,7,10,24,37)	8(2,7,8,9,17,27,38,41)	7349.34	6582.27	8224.35	5631.37
U2R	1(3)	1(3)	2(3,38)	7(2,23,31,32,35,36,39)	498.96	477.48	311.68	258.47
汇总数据	6(4,8,9,15,23,27)	11(2,4,5,11,22,26, 28,30,33,37.41)	14(2,3,8,12,16,22,31, 32,33,35,36,37,40,41)	16(2,3,4,8,12,16,17,22,23, 31,32,33,35,36,37,40)	18301.15	16459.36	16235.23	14338.53

表 4　4 种特征选择算法及所有特征的检测时间和分类准确率

入侵检测模型	分类准确率					检测时间/s				
	IPSO-TS	CMPSO	PSO-TS	PSO	所有特征	IPSO-TS	CMPSO	PSO-TS	PSO	所有特征
DoS	99.18%	95.38%	96.19%	95.37%	99.35%	19 238.52	22 881.48	22 623.93	2 3219.71	29 683.82
Probe	99.47%	96.45%	95.11%	95.54%	99.21%	941.61	1 233.48	1 333.72	1 821.40	1 743.76
R2L	98.91%	95.46%	93.88%	93.23%	98.41%	2 517.42	3 141.29	3 222.62	3 545.26	4 581.32
U2R	99.49%	95.48%	94.47%	93.78%	98.82%	0.51	0.58	0.62	0.71	0.83
汇总数据	99.53%	98.98%	93.55%	92.46%	99.45%	20 128.37	23 127.41	2 4701.14	31 682.21	38 213.46

在特征选择算法寻优特征选择数据集的特征子集，而检测阶段仅指搜寻到特征子集后分类算法 KNN 对检测样本数据集的检测时间，不包含寻优特征子集的时间。本文方法寻优到的特征子集维度更低、准确率更高，对检测样本数据集分类时处理的数据量也会更低，故可以降低分类算法的分类检测时间。

在表 1 检测样本数据集上比较上述 4 种特征选择算法后的漏报率和误报率，计算式如式(15)和式(16)所示，实验结果如表 5 所示。

$$FN\ rate = \frac{FN}{TN + FN} \tag{15}$$

$$FP\ rate = \frac{FP}{TP + FP} \tag{16}$$

从表 5 中可以得到，通过与 PSO-TS 算法、CMPSO 算法和 PSO 算法得到的特征子集作为分类样本的平均漏报率和误报率相比，本文所提 IPSO-TS 算法比 CMPSO 降低了 0.79%的平均漏报率，降低了 0.67%的平均误报率；比 PSO-TS 方法降低了 1.74%的平均漏报率，降低了 0.83%的平均误报率；比 PSO 方法降低了 1.97%的平均漏报率，降低了 1.72%的平均误报率。

本文方法以特征维数联合分类准确率构建的适应度函数作为特征筛选的评价指标，通过基于改进粒子群和禁忌搜索的二次特征选择找到全局优化特征子集。相比基于 PSO 算法或 CMPSO 算法的特征选择，本文方法首先利用改进粒子群特征选择方法找到初始优化特征子集，然后利用禁忌搜索的特征选择算法再次筛选特征子集，保证了所筛选出来的特征子集具备更高分类性能；而相比基于 PSO-TS 的特征选择方法，本文算法利用遗传算子改进了 PSO 算法，设计了适用于特征子集搜索的交叉和变异算子，实现了交叉概率和变异概率的自适应变化，有效提高了搜寻全局优化特征子集的效率。表 3~表 5 的测试结果表明，在同样的测试环境下，相比其他算法，本文方法得到的特征子集具备更低的特征维数、误报率、漏报率和检测时间，及更高的分类准确率。

5.3　实验结论

综合上述结果可以发现，本文提出的特征选择算法相比其他特征选择方法显著降低了入侵数据集的维度，提升了在线检测速度，提高了分类准确率，降低了漏报率和误报率。本文方法的主要问题是特征选择时间相比其他特征选择算法更长，但由于该时间属于离线训练阶段的处理时间，并不会增加系统的在线检测时间[19]，反而能在筛选出性能更好的特征子集后，降低系统的在线检测时延。

6　结束语

针对入侵检测中数据维度高的问题，本文所提方法通过采用遗传算子对粒子群算法进行了改进，得到了特征选择初始最优解，进一步对该解进行禁忌搜索得到了特征子集的全局优化解。该方法不仅适用于入侵检测系统，还适应于其他类似系统，例如文本分类系统，基因数据分析系统等[20]，在这些系统中，都存在对数据集进行离线特征降维和在线检测数据的需求，因此，本文方法具有一定的普适性。

参考文献：

[1] WANG C R, XU R F, LEE S J, et al. Network intrusion detection using equality constrained-optimization-based extreme learning machines[J]. Knowledge-Based Systems, 2018.

[2] 武小年, 彭小金, 杨宇洋, 等. 入侵检测中基于 SVM 的两级特征选择方法[J]. 通信学报, 2015, 36(4): 1271-1278.
WU X N, PENG X J, YANG Y Y, et al. Two-level feature selection method based on SVM in intrusion detection[J]. Chinese Journal of Communications, 2015, 36(4): 1271-1278.

[3] 张俐, 王枞. 基于最大相关最小冗余联合互信息的多标签特征选择算法[J]. 通信学报, 2018(5).
ZHANG L, WANG C. Multi-label feature selection algorithm based

表 5　　**4 种特征选择算法的漏报率和误报率**

入侵检测模型	漏报率				误报率			
	IPSO-TS	CMPSO	PSO-TS	PSO	IPSO-TS	CMPSO	PSO-TS	PSO
DoS	1.69%	2.22%	4.14%	5.29%	0.79%	1.36%	1.86%	2.20%
Probe	2.63%	3.11%	5.35%	4.34%	0.95%	2.22%	1.84%	3.55%
R2L	2.41%	2.52%	4.28%	3.70%	1.41%	2.38%	2.56%	3.88%
U2R	1.43%	3.42%	2.52%	2.95%	0.83%	1.32%	1.29%	2.82%
汇总数据	1.14%	2.02%	1.69%	2.85%	1.79%	1.85%	2.38%	1.92%

on maximum correlation minimum redundant joint mutual information[J]. Transactions of Communications, 2018(5).

[4] VIEIRA S M, MENDONCA L F, FARINHA G J, et al. Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients[J]. Applied Soft Computing, 2013, 13(8): 3494-3504.

[5] XUE B, ZHANG M, BROWNE W N, et al.A survey on evolutionary computation approaches to feature selection[J].IEEE Transactions on Evolutionary Computation,2016,20(4):606-626.

[6] GHAMISI P, BENEDIKTSSON J A. Feature selection based on hybridization of genetic algorithm and particle swarm optimization[J]. IEEE on Geoscience and Remote Sensing Letters, 2015, 12(2): 309-313.

[7] 董跃华, 刘力. 基于自适应改进粒子群优化的数据离散化算法[J]. 计算机应用, 2016, 36(1):188-193.
DONG Y H, LIU L. Data discretization algorithm based on adaptive improved particle swarm optimization[J]. Journal of Computer Applications, 2016, 36(1): 188-193.

[8] TRAN B, XUE B, ZHANG M. A new representation in PSO for discretization-based feature selection[J]. IEEE Transactions on Cybernetics, 2017, PP(99):1-14.

[9] NGUYEN H B, XUE B, ANDREAE P, et al. Particle swarm optimisation with genetic operators for feature selection[C]. Evolutionary Computation. IEEE, 2017: 286-293.

[10] 翟俊海, 刘博, 张素芳. 基于粗糙集相对分类信息熵和粒子群优化的特征选择方法[J]. 智能系统学报, 2017, 12(3):397-404.
ZHAI J H, LIU B, ZHANG S F. Feature selection based on rough classification relative classification information entropy and particle swarm optimization[J]. Journal of Intelligent Systems, 2017, 12(3): 397-404.

[11] 刘杨,田学锋,詹志辉.粒子群优化算法惯量权重控制方法的研究[J]. 南京大学学报:自然科学版,2011,47(4): 364-371.
LIU Y, TIAN X F, ZHAN Z H. Study on inertia weight control method based on particle swarm optimization algorithm[J].Journal of Nanjing University: Nature Science, 2011,47(4): 364-371.

[12] ZHANG Q, XUE S. An improved multi-objective particle swarm optimization algorithm[J]. Mathematical Problems in Engineering, 2017, 28(7):482-490.

[13] BHARTI K K, SINGH P K. Opposition chaotic fitness mutation based adaptive inertia weight BPSO for feature selection in text clustering[J].Applied Soft Computing, 2016, 43:20-34.

[14] LAI X, YUE D, HAO J K, et al. Solution-based tabu search for the maximum min-sum dispersion problem[J]. Information Sciences, 2018.

[15] KUO S Y, CHOU Y H. Entanglement-enhanced quantum-inspired tabu search algorithm for function optimization[J]. IEEE Access, 2017, PP(99):1-1.

[16] HOU N, HE F, CHEN Y. An adaptive neighborhood taboo search on GPU for Hardware/Software Co-design[C]//International Conference on Computer Supported Cooperative Work in Design. 2016: 239-244.

[17] JANARTHANAN T, ZARGARI S. Feature selection in UNSW-NB15 and KDDCUP'99 datasets[C]// International Symposium on Industrial Electronics. 2017:1881-1886.

[18] ZHANG S, LI X, ZONG M, et al. Learning k, for kNN classification[J]. ACM Transactions on Intelligent Systems and Technology, 2017, 8(3):43.

[19] DAS S, LIU Y, ZHANG W, et al. Semantics-based online malware detection: towards efficient real-time protection against malware[J]. IEEE Transactions on Information Forensics and Security, 2017, 11(2): 289-302.

[20] ZHANG Y, YANG A, XIONG C, et al. Feature selection using data envelopment analysis[J]. Knowledge-Based Systems, 2014, 64(64): 70-80.

[作者简介]

张震（1985−），男，山东济宁人，博士，国家数字交换系统工程技术研究中心讲师，主要研究方向为网络测量、网络管理。

魏鹏（1994−），男，湖南衡阳人，国家数字交换系统工程技术研究中心硕士生，主要研究方向为新型网络体系结构。

李玉峰（1976−），男，山东烟台人，博士，国家数字交换系统工程技术研究中心副教授，主要研究方向为宽带信息网络、高速路由器核心技术。

兰巨龙（1962−），男，河北张北人，博士，国家数字交换系统工程技术研究中心教授、博士生导师，主要研究方向为宽带信息网络。

徐萍（1983−），女，江西永新人，中国人民解放军战略支援部队信息工程大学讲师，主要研究方向为信息素质教育、信息资源建设。

陈博（1989−），男，河南商丘人，国家数字交换系统工程技术研究中心博士生、讲师，主要研究方向为网络安全。